SPEECH PATTERNS IN VIDEO-MEDIATED CONVERSATIONS

*Abigail J. Sellen**

Computer Systems Research Institute
University of Toronto
6 King's College Rd.
Toronto, Ontario
M5S 1A1 Canada

ABSTRACT

This paper reports on the first of a series of analyses aimed at comparing same room and video-mediated conversations for multiparty meetings. This study compared patterns of spontaneous speech for same room versus two video-mediated conversations. One video system used a single camera, monitor and speaker, and a picture-in-a-picture device to display multiple people on one screen. The other system used multiple cameras, monitors, and speakers in order to support directional gaze cues and selective listening. Differences were found between same room and video-mediated conversations in terms of floor control and amount of simultaneous speech. While no differences were found between the video systems in terms of objective speech measures, other important differences are suggested and discussed.

KEYWORDS: CSCW, videoconferencing, conversation patterns.

INTRODUCTION

People meet for a variety of reasons: to discuss and share ideas, to argue and make decisions, to plan, and to socialize. Video and audio technology has obvious potential for bringing people together at remote locations. Cameras and microphones provide electronic eyes and ears; monitors and speakers deliver visual and auditory information. Combined with computer supported groupware such as electronic whiteboards, all the right ingredients for a simulated face-to-face meeting seem to be in place.

There are nonetheless important differences between video and face-to-face (or same room) meetings. Some of these are rather obvious. Unlike eyes, cameras have a fixed field of view and usually cannot be controlled by the viewer. Failure to make eye contact tends also to be a problem because of separation of camera and monitor. In video-mediated meetings, the principle of reciprocity does not always hold (i.e., if I can see you, you can see me). There is no concept of a negotiated mutual distance between speakers, and speakers have no sense of how their voices are perceived by listeners. Other differences are more subtle and harder to define, such as the relative impotence of gestures and gaze in securing another's attention through video [13], and the feeling of being "distanced" from others.

Many of these problems are compounded when one is restricted to a single camera and monitor in order to converse with multiple parties. One way of supporting multiparty conversations is to use a "picture-in-a-picture" device which divides the screen into quadrants with one participant occupying each quadrant. However, when multiple participants occupy a single screen, participants are limited in their ability to: 1) direct their gaze to various participants; 2) establish eye contact with other participants; 2) be aware of who, if anyone, is visually attending to them; 3) selectively listen to different, parallel conversations; 4) make aside comments to other participants; and 5) hold parallel conversations.

This experiment was conducted in order to compare same room conversations with video-mediated conversations, and also to compare conversational behavior in two video systems. One video system uses the picture-in-a-picture (PIP) approach and thus suffers from the limitations listed above. The other, called *Hydra*, was designed specifically to support these abilities.

A series of analyses are planned for the data collected in this study. This paper reports on the first analysis — an examination of the gross structure of conversation for each of the three conditions. The general question of interest is how video-mediation affects conversational structure in terms of the on-off patterns of speech. A more specific question is whether the properties of the Hydra video system are sufficiently different from the PIP approach to affect speech patterns. For example, head turning and gaze cues are thought to be important in regulating the flow of conversation. Since Hydra is intended to support these kinds of cues, this may be reflected in objective measures of speech. These issues are addressed in the context of discussions involving four people.

While there are many interesting theoretical issues that arise, there are also practical issues motivating this work. This study is being carried out within the larger context of

*Author is now at Rank Xerox Cambridge EuroPARC, and the MRC Applied Psychology Unit, Cambridge, UK. E-mail: sellen@europarc.xerox.com

the CAVECAT project [16], a project which is exploring a variety of issues in technology supported cooperative work. In particular, this study is part of a more comprehensive design effort examining different ways of supporting multiparty videoconferencing.

The "Hydra" System

Hydra is a system which uses multiple cameras, monitors and speakers to support multiparty videoconferencing [19]. Hydra simulates a 4-way round-table meeting by placing a camera, monitor and speaker in the place that would otherwise be held by each remote participant. Using this technique, each person is presented with a unique view of each remote participant, and that view, and its accompanying voice, emanates from a distinct location in space. Figure 1 shows Hydra in use in a four-way conversation.

The fact that each participant is represented by a separate camera/monitor pair means that gazing toward someone is effectively conveyed. In other words, when person A turns to look at person B, B is able to see A turn to look towards B's camera. Looking away and gazing at someone else is also conveyed, and the direction of head turning indicates who is being looked at. Furthermore, because the voices come from distinct locations, one is able to selectively attend to different speakers who may be speaking simultaneously.

Audio and video connections for Hydra are configured by software which ensures that a consistent "around the table" mapping is made for each person. In other words, the switching network ensures that if person A appears in the center unit for person B, then B appears in the center unit for person A. Similarly, if person C appears to person A's right, then person C appears to person B's left, and so on. In this way, head turning and gaze cues deliver consistent and meaningful information.

Gaze and the Regulation of Conversation

In this experiment, whether directional gaze cues were present in the conversation was one factor of interest. It is estimated that 60 percent of conversation involves gaze and 30 percent involves mutual gaze [1]. Gaze serves at least five functions [2,15]: to regulate the flow of conversation; to provide feedback on how the communication is being perceived by the listener; to communicate emotions; to communicate the nature of the interpersonal relationship; and to avoid excess information input. Video systems which fail to support gaze and mutual gaze may affect any of these five functions.

One effect which may reveal itself in patterns of conversation is in the regulation of conversation, or how floor control is passed from speaker to speaker. There are a variety of different cues which are used to coordinate turn-taking such as intonation, paralanguage, body motion, and syntax [10]. Among these, gaze and head turning have been well established as being used to keep the floor, to take the floor, to avoid taking the floor, and to suggest who should speak next [2,10,11,12,15]. Kendon [15] found that gaze by a speaker at a listener increases just before ending a long utterance, and that when there is no such terminal gaze, there was more likely to be a pause before switching speakers. In general, a speaker will tend to look away at the beginning of a turn and then terminate the turn with a sustained gaze, usually at presumptive next speaker. A speaker wishing to hold the floor at a pause point will look away from the listener.

Figure 1. A user is seated in front of three Hydra units. Each Hydra unit contains a video monitor, camera, and loudspeaker.

As Short, Williams and Christie [21] have noted, reintroducing the visual channel via conventional video systems may exacerbate problems in regulating conversation. Not only do head turning and directional gaze cues tend to be eliminated, asymmetry may also be an important aspect of the problem. For example, one participant may believe that they are making eye contact, but this is not perceived by the other participant. Similarly, participants from time to time will look at the camera, and this may be interpreted as a signal by the receiver of the look. There is some empirical evidence to support the fact that asymmetry can be problematic. Argyle, Lalljee, and Cook [4] found that asymmetry in the amount of visibility between conversants led to greater effects in terms of pauselength and interruptions than symmetrical lessening of visual cues.

The Experiment

In this study four-person groups were used. Number of participants is an important consideration. Dyads are typically the basis of research on conversational structure, in part because they are simpler to study. However, as soon as a third party is introduced, "next turn" is no longer guaranteed to the non-current speaker. Further, three party conversations are notably different from four-party conversations in that four people provides for the possibility of two different ongoing conversations. Four people in a discussion also means that it is potentially more difficult to gain the floor. Thus it was hoped that using a larger group would accentuate differences between conditions in terms of regulation of conversation.

There are few studies which have objectively measured patterns of spontaneous speech across media. Studies of dyadic conversations in audio-only conditions (e.g., telephone) have found that when one takes away visual cues, there tend to be fewer interruptions [7,18], shorter periods of simultaneous speech [14,18], and longer utterances [18]. However, there is some conflicting evidence. Argyle, Lalljee and Cook [4] found *more* interruptions during dyadic conversations when visual cues were reduced. The findings with regard to pauses are also inconsistent. Argyle, Lalljee and Cook [4] found longer pauses when visual cues are reduced. Cook and Lalljee [7] found no difference between media for length of pauses. Jaffe and Feldstein [14] found slightly shorter pauses within utterances and between switching speakers in a no-vision condition, for mixed sex pairs only.

Only Cohen [6] has objectively measured conversational parameters for groups of more than two people in the context of face-to-face versus video conditions. Cohen compared a face-to-face condition with a meeting using Bell's Picturephone Meeting Service (a voice switched system consisting of 3 cameras in each of two rooms, with three to four people per room). Objective measures of conversational structure showed that the face-to-face condition resulted in more speaker turns and more simultaneous speech than the Picturephone Meeting Service. It seems then, that these results are consistent with most of the studies of dyads, and that video conditions may have similar effects on conversation patterns as audio-only conditions.

Some have taken these findings to mean that technology-mediated conversations are *better* synchronized than face-to-face meetings. Technology mediated conversants experience fewer interruptions, and turn-taking appears to be more orderly. Regulation of turns is obviously not wholly dependent on face-to-face visual cues. The audio channel also carries synchronization cues and perhaps compensates for the loss or attenuation of visual cues. However, what has been called more "orderly" conversation may in fact reflect a reluctance on the part of listeners to interject and try to seize the floor when visual cues are attenuated. Rutter [17] showed that no-vision discussions are perceived to be less spontaneous, more formal and more socially distant than face-to-face discussions.

Experimental Hypotheses

Because Hydra is designed to simulate a four-way meeting using video surrogates, the overriding expectation was that Hydra would tend to produce conversational patterns more similar to same room conversations than a PIP approach. The PIP approach not only fails to support selective gaze and listening, but it is designed so that a viewer sees themselves in addition to the other three people. This design feature also is unlike a face-to-face situation, and thus was thought to contribute to its "unnaturalness", perhaps to the extent that it affects the structure of conversation. With this in mind, and by extrapolating from the existing literature, the following hypotheses were put forth:

H1. Same room conversations will result in the highest number of turns per session. The fewest will occur in the PIP condition.

H2. The average duration of turns will be shortest in the same room condition, and longest in the PIP condition.

Hypotheses 1 and 2 are based on Cohen's [6] finding that there were almost twice as many speaker switches in a face-to-face meeting than in a Picturephone meeting. If this finding holds for other kinds of video systems, we would expect more frequent and shorter turns in a conversation, all else being equal.

H3. There will be the most unequal distribution of turns among speakers in the PIP condition, and the most equal distribution in the same room condition.

This hypothesis is based on the assumption that if it is more difficult to switch speakers (i.e. in the video conditions), it will be done less often. Thus in the video conditions, and especially in the PIP condition, dominant speakers will dominate more, and non-dominant speakers will attempt to take the floor less often.

H4. There will be more simultaneous speech in same room condition than in the two video conditions.

The Hydra system will produce more simultaneous speech than the PIP condition.

This hypothesis is based on Cohen's [6] results which found more simultaneous speech in face-to-face meetings than in Picturephone meetings. Cohen concluded that face-to-face meetings were thus less polite, less orderly, and more interactive. If Hydra is more like a same room meeting than the PIP system, Hydra conversations should be more interactive, and less orderly than PIP conversations.

No specific hypothesis regarding time between speaker switches was put forth. As discussed previously, the data regarding the effect of visual cues on switch pauses are inconsistent. However, if perfect coordination between speakers means minimizing interruptions and minimizing pauses, then a zero switching time is ideal. The average switch pause is typically in the range of .62 to .77 seconds [14]. As coordination gets worse, we would expect longer switch times to occur. However, we might also expect more overlapping speech during speaker switches. These overlaps are typically not examined separately but are classified as simultaneous speech. One alternative is to conceptualize switching time as a single metric which is sometimes negative (an overlap) and which is sometimes positive (a switch pause). The effect of video-mediation on this measure remains to be explored.

METHOD

Subjects

Twelve groups of four adults participated: 15 women and 33 men. With only a couple of exceptions, none of the subjects knew each other previously.

Task and Experimental Design

Each group was asked to participate in a set of three informal debates lasting approximately sixteen minutes each. Subjects were randomly divided into teams of two, and each team was randomly assigned either to the "Pro" or "Con" side of the issue. Three different topics were introduced with the help of one or two short newspaper clippings. The topics were: the right to smoke in public, mandatory drug testing, and censorship in the news. Each group discussed all three topics, one in each condition. Teams remained the same for all three topics and topics were counterbalanced across conditions.

This was a simple one-factor repeated measures design, comparing performance in three conditions: Same Room, Picture-in-a-Picture video system, and the Hydra video system. Order of condition was counterbalanced using a Latin square design.

Experimental Conditions and Apparatus

The three conditions are described below. Both audio and video records of each conversation were made using a video camera and a VCR located in a separate control room. In addition, specialized speech tracking equipment (also described below) was used in order to record on-off patterns of speech in the conversation.

Same Room Condition. In this condition, all four subjects met in the same room around a table. A video camera was set up in one corner of the room and the video output was channeled through coaxial cable to a VHS videorecorder in the experimental control room. In addition, each subject wore a headset microphone. This audio output was also fed through coaxial cable to the experimental control room. There, it went both to a mixer where all four voices were laid down on the audio track of the video cassette, and to the speech tracking equipment, also located in the control room.

Picture-in-a-Picture (PIP) Condition. Each subject was seated in separate room outfitted with a color video monitor, video camera, a speaker, and a headset microphone. The camera was mounted on top of each monitor and the speaker was located immediately adjacent to each monitor. A video board allowed the display of four composite images as illustrated in Figure 2. This configuration allowed each participant to see the other three participants as well as an image of themselves. Each subject saw exactly the same configuration of images as the other subjects.

As in the Same Room condition, video and audio recordings were made of each conversation. Also as before, the audio output from each microphone was mixed and laid down on the videotape, in addition to being sent to the speech tracking equipment.

Hydra Condition. The Hydra system was set up in each of the same rooms used in the PIP condition. Each of the three Hydra units was constructed from a Sony Watchman color monitor (8 cm diagonal), a black and white camera from a Radio Shack surveillance unit mounted 4.5 cm below the screen, and a speaker mounted just below the camera, also from a Sony Watchman. Each unit tilts back and forth for best viewing position. In the other three rooms, simulated Hydra units had to be used due to budget constraints. In these rooms, three Radio Shack black and white monitors were used (12 cm in diameter), along with two black and white Radio Shack surveillance cameras, and the color camera used in the PIP condition. The color camera was used to feed the prototype Hydra units in order to take advantage of the color monitors in those units. Each camera was mounted directly on top of each monitor. In addition, each camera/monitor pair was mounted directly on top of a speaker. In all cases, the Hydra or simulated Hydra units were located 15 cm apart on the desk top, and set back 38 cm from the edge of the desk.

Figure 2. A meeting using the picture-in-a-picture (PIP) device. Each person sees the other three people in addition to themselves, on one screen. All participants see the same image.

Speech Tracking System
The conversion of speech into digital on/off patterns was accomplished by obtaining audio output from each of the four subjects using unidirectional dynamic, headset microphones. Each microphone output controlled its own externally keyed audio noise gate. When a subject spoke louder then a preset threshold, the corresponding audio noise gate would open, allowing a fixed pitch generated by a Yamaha TX802 synthesizer to pass through. When a subject fell silent the gate would close and cutting off the pitch. Each of the output signals from the four noise gates were fed into four input channels of an IVL Pitchrider 7000 Mark II pitch tracking device. The pitch tracker converted the pitch on/off signals into digital on/off signals and send them, via a MIDI connection, to a Macintosh II computer. These on/off events were stored in the computer and each event was time stamped with SMPTE time code. This time code was simultaneously laid down on the videotape so speaker events could be later synchronized when playing back the videotape.

Procedure
On arrival, subjects on the same team were introduced to each other and given approximately 15 minutes to get acquainted while completing the experimental consent forms. Following this, they were introduced to the members of the other team and were instructed to read the first topic for debate. They were then placed in separate rooms (in the case of the PIP or Hydra conditions) or in the same room, and were instructed on wearing the headset microphones. All three conditions used a similar procedure. Subjects discussed the prescribed topic for 16 minutes, and then were asked to complete a questionnaire about the conversation they had just experienced, independently of each other.

RESULTS

Analysis of Speech Data
Each 16 minute conversation was checked for accuracy against the videotape data, edited where necessary, and coded using specialized software designed for this purpose. Despite the impressive accuracy of the speech tracking system, some sporadic crosstalk did occur from time to time which had to be deleted. In addition, 200 msec pauses were filled in, in order to account for stop consonants (a procedure also used by Brady, [5]). Laughter and also backchannel responses were coded so as to differentiate these data from speaker turns or attempts to take turns. Backchannel responses are vocalizations such as "mmm-hmm", often used to show attentiveness, which do not constitute turns or attempts to take turns [10].

Definitions
The data were analyzed using definitions taken both from Jaffe and Feldstein [14] and Dabbs and Ruback [8,9], and then modified slightly. Dabbs and Ruback's scheme is an extension of that of Jaffe and Feldstein to better account for groups larger than dyads.

The following definitions were used:

Turn. A turn consists of the sequence of talkspurts and pauses by a speaker who "has the floor". A speaker gains the floor when they begin speaking to the exclusion of everyone else and when they are not interrupted by anyone else for at least 1.5 seconds.[1] The duration of a turn begins with the first unilateral sound, and ends when another

[1] Without this criterion, even the shortest unilateral sound would be designated as a turn. 1.5 seconds was chosen because this is estimated to be the mean duration of a phonemic clause and there is evidence for the phonemic clause as a basic unit in the encoding and decoding of speech [14].

individual turn or a "group turn" begins (see below). Note that turns therefore include periods of mutual silence at the end of utterances, when no one else has yet taken the floor.

Group Turn. Using Dabbs and Ruback's [8] definition: "A group turn begins the moment an individual turn taker has fallen silent and two or more others are speaking together; the group turn ends the moment any individual is again speaking alone" (p. 519). Dabbs and Ruback proposed the group turn to cover instances where individual turn takers are effectively "drowned out" by the group.

Speaker Switch. A speaker switch occurs whenever one person or group loses the floor, and another person or group gains it.

Switch Time. Switch time consists of *switching pauses* and *overlaps*. A switching pause is a period of mutual silence bounded by different turn takers (individuals or groups). Unlike existing definitions, I also include as a related measure the concept of overlap. An overlap is a period of simultaneous speech immediately before and leading to the person who utters it taking a turn. The two measures can be conceptualized as a single continuous parameter which measures the relationship between one person ending a turn and another starting. A negative switch time is thus an overlap, while a positive switch time is a switch pause.

Simultaneous Speech. Simultaneous speech is speech by one or more speakers who do not have the floor. I further distinguish between overlaps and simultaneous speech which does not lead to a speaker switch. Simultaneous speech which does not precede a speaker switch is called *non-interruptive simultaneous speech* . Overlaps are synonymous with *interruptive simultaneous speech.*

Turn Analysis

The number, duration and distribution of speaker turns is shown in Table 1. Statistical tests using one-tailed analyses of variance at the .05 level showed:

1. No difference across conditions in the number of individual turns per session.

2. No difference across conditions in the mean duration of individual turns.

3. No difference across conditions in the number of group turns per session.

4. No difference across conditions in the distribution of turns among speakers.

Turn distribution among speakers was calculated after Dabbs and Ruback [8] who used Shannon and Weaver's [20] equation for calculating information (in information theory terms). This equation defining H, or amount of information, is essentially a way of calculating the average amount of uncertainty about who has the floor at any given time. H is defined by:

$$H = -\Sigma p_i \log (p_i)$$

If one person talks all the time, H is equal to its minimum value of zero (no uncertainty). If all four people hold the floor for an equal number of turns, H is equal to 2, its maximum value.

Table 1. Average number and duration of individual turns, average number of group turns per session, and distribution of speaker turns. Standard deviations are shown in brackets.

	Overall F Tests	Same Room	PIP	Hydra
1. Number of Turns per Session	not sig.	62.6 (17.4)	64.1 (19.5)	68.7 (24.25)
2. Turn Duration (sec)	not sig.	15.92 (3.66)	16.46 (6.66)	16.62 (10.31)
3. Number of Group Turns per Session	not sig.	3.8 (3.8)	4.6 (5.0)	3.8 (6.9)
4. Distribution of Turns (H value)	not sig.	1.83 (.10)	1.82 (.17)	1.83 (.17)

Simultaneous Speech Analysis

Table 2 presents the data summary for simultaneous speech and switching measures. Analyses of variance (two-tailed tests at the .05 level) showed:

5. *Percentage of time one person spoke* did not differ significantly across conditions. However, this measure almost reached significance ($F(2,22) = 2.87$, $p < .078$). Percent time of one person talking was based on the summation of all time intervals during which one person only spoke, expressed as a percentage of total session time (960 secs).

6. *Percentage of simultaneous speech* was significantly different across conditions. Means comparisons showed Same Room conversations to contain more simultaneous speech than the video conditions ($F(1,22) = 6.78$, $p < .016$), but showed no difference between the two kinds of video conditions. Percent of simultaneous speech refers to the proportion of time during which two, three or four people were speaking simultaneously.

7. No difference was found in the total amount of *non-interruptive simultaneous speech* across conditions, although the differences were close to significant ($F(2,22) = 2.56$, $p < .10$). Amount of non-interruptive simultaneous speech is the sum of all simultaneous speech events not leading to a speaker switch.

8. *Amount of interruptive simultaneous speech* was found to differ across conditions. Means comparisons showed that Same Room conversations gave rise to more interruptive simultaneous speech than the video conditions ($F(1,22) = 7.04$, $p < .015$), with no difference between video conditions. Amount of interruptive simultaneous speech is the sum of all simultaneous speech events which result in the interrupting speaker taking the floor.

9. *Percent of simultaneous speech taking control* did not differ across conditions. Percent of simultaneous speech taking control is the percentage of simultaneous speech which is interruptive (as opposed to that which does not eventually take the floor).

10. *Percent of speaker switches consisting of overlaps* (as opposed to pauses) did differ across conditions. More speaker switches consisted of overlaps in the Same Room conditions than the video conditions ($F(1,22) = 7.85$, $p < .01$), with no difference between video conditions. Percent of overlaps in speaker switches calculates what percentage of speaker switches takes place with a negative switching time (an overlap), rather than a switch pause.

11. *Switching time* was significantly different across conditions. The Same Room condition gave rise to a mean switch overlap, while the video conditions gave rise to a positive switch time value, or switch pause. The difference between Same Room and video conditions was significant ($F(1,22) = 27.67$, $p < .0001$), but no difference between video conditions was found. Switching time is an average of switch pauses (positive values), and overlaps (negative values).

Table 2. Summary statistics for simultaneous speech, percent of simultaneous speech taking control of the conversation, and switching time. Standard deviations shown in brackets.

	Overall F Tests	Same Room	PIP	Hydra
5. % Time One Person Talking	(*) $p < .078$	71.3 (3.7)	72.9 (4.1)	74.7 (5.5)
6. % Simultaneous Speech	* $p < .033$	9.7 (7.4)	7.1 (6.9)	5.4 (7.1)
7. Am't of Non-Interruptive Simultaneous Speech (sec)	(*) $p < .10$	98.1 (86.9)	72.1 (67.7)	57.8 (80.0)
8. Am't of Interruptive Simultaneous Speech (sec)	* $p < .038$	56.6 (39.2)	40.1 (39.5)	33.6 (39.7)
9. % Simultaneous Speech Taking Control	not sig.	38.5 (6.6)	34.4 (5.9)	41.6 (15.6)
10. % Overlaps in Speaker Switches	* $p < .029$	54.1 (18.9)	46.3 (24.1)	43.5 (21.1)
11. Switching Time (sec)	** $p < .0001$	-.46 (.66)	.04 (.79)	.25 (.67)

Questionnaire Data
The mean scores from the questionnaires averaged across 48 subjects are shown in Table 3. Analysis of variance tests found four statistically significant results:

12. Subjects rated the Same Room meeting as allowing them to better take control of the conversation than both video conditions ($F(1,47) = 10.59$, $p < .002$). There was no difference between video conditions.

13. Subjects rated the Same Room conversation as being more interactive than either video condition ($F(1,47) = 5.65$, $p < .022$). There was no difference between video conditions.

14. Subjects rated the Same Room conversation as allowing them to selectively attend to one person at a time most easily ($F(1,47) = 8.73$, $p < .005$). While the Hydra video system gave rise to a higher overall mean than the PIP system, this difference did not reach significance.

15. Subjects rated the Same Room condition the best for knowing when others were listening or attending to them. This was rated significantly better than the Hydra system ($F(1,47) = 22.18$, $p < .0001$), which was rated significantly better than the PIP system ($F(1,47) = 12.13$, $p < .001$).

DISCUSSION

Turn Frequency, Duration, and Distribution
Mediating conversations with video technology appeared to have no discernable effects on the number of turns taken per session, the average length of those turns, or on the distribution of turns among speakers. These results were unexpected, especially considering previous research which generally finds that audio-only conditions, and in one case, a video-mediated condition [6], tend to increase turn length relative to face-to-face conversations. In light of the lack of differences between same room conversations and video-mediated conversations, it is perhaps not surprising that no difference was found between the two video conditions on these measures.

Table 3. The mean scores for each of the nine questions administered in the questionnaires averaged over 48 subjects. A score of 7 represents "Strongly Agree", while a score of 1 represents "Strongly Disagree".

Question	Same Room	PIP	Hydra
I was able to talk and express myself freely.	6.1 (0.9)	5.8 (1.4)	5.7 (1.4)
I was able to take control of the conversation when I wanted to. **($p < .002$)**	6.0 (1.0)	5.4 (1.5)	5.5 (1.4)
There were too many inappropriate interruptions.	2.4 (1.5)	2.4 (1.4)	2.2 (1.4)
This was an unnatural conversation.	2.7 (1.7)	3.1 (1.7)	3.0 (1.6)
The conversation seemed highly interactive. **($p < .006$)**	5.9 (1.3)	5.4 (1.3)	5.3 (1.2)
There were many unnatural and uncomfortable pauses.	2.5 (1.7)	2.3 (1.1)	2.6 (1.4)
I could selectively attend to one person at a time. **($p < .0001$)**	6.1 (1.1)	4.8 (1.8)	5.3 (1.8)
I knew when people were listening or paying attention to me. **($p < .0001$)**	6.3 (0.9)	4.3 (1.9)	5.3 (1.5)
I found it difficult to keep track of the conversation. **($p < .09$)**	2.0 (1.5)	2.6 (1.7)	2.2 (1.2)

The discrepancy between Cohen's [6] results and these results may be due to the design of the Picturephone Meeting Service she used. Picturephone is a voice activated system which, in her study, switched between six different cameras depending on who in the group was talking. This meant that the whole group could never be viewed simultaneously. She also introduced a 705 msec audio and video transmission delay in order to simulate round-trip satellite conditions. These two factors could well account for differences in turn length and frequency, since this design would presumably more radically reduce the effectiveness of both verbal and visual cues to regulate turn-taking behavior.

Perhaps the results of this study with respect to turn frequency, duration, and distribution speak to the success of both the PIP and Hydra approach in preserving the structure of the conversation, at least at this level. The results of the questionnaire confirm that subjects did not feel that any of the three different situations was especially unnatural or uncomfortable. In both systems, and unlike the Picturephone system, participants are visually available all the time. Thus each person can monitor all other members of the group whether they are speaking or not and non-verbal signals for turn-taking can be perceived. Showing that this factor alone accounts for differences between Cohen's results and these results would require running a voice switched video condition with no audio or video transmission delay.

A final point to note is that the groups were highly variable in overall amount of talking, amount of simultaneous speech, and distribution of turns among speakers. Pronounced between group differences can be contrasted with relatively stable group characteristics across conditions. This emphasizes the importance of using within-group designs for this kind of study.

Simultaneous Speech and Floor Control

Subjects did feel it was more difficult to take control of the conversation in the video conditions than in the Same Room condition (as evidenced by the questionnaire data). Nonetheless, this difficulty was not reflected in the distribution of turns among speakers, as might be expected. Where differences do emerge, however, is in the amount of simultaneous speech that occurred and in the time between switching speakers.

A lower percentage of time was occupied by one speaker talking, and a higher percentage of time was occupied by simultaneous speech in the Same Room condition relative to the two video conditions. This result is in line with previous findings for audio-only and video-mediated conversation, although some researchers have found more interruptions when visual cues are reduced [4].

A more informative analysis may come from asking what function simultaneous speech serves, or what it may indicate. On the one hand, simultaneous speech may be taken to indicate a problem in floor control. Participants may mistime their bids for floor control, or may bid for the floor and fail. Studies which label simultaneous speech as "interruptions" make this tacit assumption. On the other hand, simultaneity may also be taken to be an indication of the degree of interactivity and spontaneity of the conversation. Conversations which have more simultaneous speech may be due to participants who feel more engaged in the conversation, and are more willing to attempt to take the floor.

Rather than attaching a value judgement to simultaneous speech, it may be more useful to distinguish between simultaneous speech which gains control of the floor versus that which does not. One can then discover how often attempts at floor control occur, and how often they are successful. Most existing studies do not make this distinction.

Video conversations gave rise to less non-interruptive simultaneous speech (although not significantly less), and less interruptive simultaneous speech overall. In addition, the Hydra system gave rise to less simultaneous speech of both types than the PIP system, although this difference was not significant. What this may indicate is a reluctance on the part of conversants to attempt to take the floor in video-mediated conversations. This is in line with many subjects' spontaneous comments. Many reported feeling "distanced" by the video systems, and less a part of the conversation. Perhaps they felt that bids for floor control would be less effective in video-mediated conversations.

The actual effectiveness of bidding for the floor while someone else is talking can be estimated by calculating the percentage of simultaneous speech that gains the floor. As is shown in Table 2, simultaneous speech was successful in gaining the floor about 34 to 42 percent of the time, and the differences across conditions was not significant. Thus, there was no *real* difference in the probability of bids for the floor being effective in Same Room versus video conditions.

If subjects were more reluctant to bid for the floor in the video conditions, and bidding was equally effective, why would this not result in fewer speaker switches in the video conditions? The answer may lie in the fact that speaker switching in the Same Room condition was more likely to occur with an overlap between speaker turns than a pause. Speaker switching in the video conditions, on the other hand, was more likely to occur with a brief pause. The analysis of switching time confirms this finding. Switching time in the Same Room condition was -.46 seconds on average, while mean switching time in the video conditions was a positive value (.04 sec for the PIP condition, and .25 sec for Hydra). It is as if conversants in video-mediated conversations were more opportunistic or polite, waiting for a pause or for a speaker to finish before attempting to take the floor. This theory is speculation at this point, however. A clearer picture will likely emerge after a more thorough analysis of the videotape data.

PIP versus Hydra Systems
Contrary to expectation, there were no differences between the two video systems in terms of objective measures of on-off patterns of speech. However, both the questionnaire data and informal discussions with subjects after each experimental session confirmed that subjects did notice differences between the systems, and most had strong opinions on which system they preferred.

The majority of subjects preferred the Hydra system. Reasons given included the fact that they could selectively attend to people, and could tell when people were attending to them. Another frequent comment was that they liked the multiple sources of audio in the Hydra system, and that this helped them keep track of one thread of the conversation when people talked simultaneously. The questionnaire data confirm that keeping track of the conversation in the PIP condition was the most difficult. Thus, it is reasonable to conclude that Hydra was successful in facilitating selective listening and selective gaze, in line with the original intent behind its design.

A preliminary analysis of the videotape data also confirms that Hydra was successful in affording aside and parallel conversations. Separate conversational threads occurred concurrently a total of four times in the Hydra condition, and three times in the Same Room condition, but never in the PIP condition. Therefore, even though no differences appeared in the structural analysis of speech, it seems likely that an in-depth analysis of the videotapes will reveal differences that do exist between these two systems.

Why the selective gaze and headturning cues did not affect the structure of the conversation is an interesting issue. Head turning and directional gaze could be readily observed in the Hydra conversations. However video-mediation may render these kinds of cues ineffective for their recipients. As Heath and Luff [13] have pointed out, movements in the periphery which appear on a screen lose their power to attract attention. Presumably this is even more of a problem for small screens. Speakers may also face difficulties in knowing how their gestures are received. Indeed, many subjects commented that they wanted a mirror to see how they were framed from the point of view of others. Thus even though Hydra is designed to support directional gaze cues, video mediation may nonetheless detract from the ability of such cues to affect behavior.

Finally, about one third of the subjects preferred the PIP system to the Hydra system. It was interesting to find that most of these subjects commented that they enjoyed having all of the participants on one screen because it meant that head turning was *not* necessary. Some subjects said they liked to see themselves to know how they were seen by others, even though this could sometimes be distracting. One subject commented that seeing herself on the same screen as the others made her feel more part of the group, and said that otherwise she would have felt quite distanced from them. Thus, simulating aspects of face-to-face situations need not always provide the correct design solutions. The PIP system seems to overcome some of the problems inherent in video mediation, such as the feeling of being distanced and its inherent lack of reciprocity.

CONCLUSIONS

This paper provides some statistics on differences between same room and video-mediated conversations for multiparty conversations. However, some unexpected similarities were also discovered. Videoconferencing did not seem to have much effect on how often people spoke, or for how long, or on the patterns of distribution of turns among group members. Both video systems used in this experiment have the characteristic that participants are visually present all the time and no one person "owns" the audio channel. This may account for the lack of drastic differences between conditions. Other kinds of systems such as voice-activated video switching systems have the characteristic that only the current speaker is displayed to the other participants. This aspect of design may result in much larger effects on conversational structure. We currently have such a system in place and are running a second study to test this assumption.

Very few, if any, studies exist which compare objective measures of conversation for different kinds of video systems. This paper provides some of those statistics, and has also shown that such measures may be relatively insensitive to more subtle but nonetheless important aspects of videoconferencing system design. One factor which may have downplayed any differences was the small monitors used in the Hydra design. Because image size was so small, this may have decreased the effectiveness of directional gaze cues in peripheral vision.

Despite this finding, there is every indication that significant differences between the two systems examined here do exist. Both the ability to selectively attend to different audio streams, and to different video images appeared to be successful in making aside conversations possible. In addition, subjects commented that multiple speakers made it easier to follow the conversation. These are clearly important aspects of design, and a more in-depth, qualitative analysis of the videotapes is planned to explore them further.

ACKNOWLEDGEMENTS

I owe a great deal of thanks to four people who became entangled in a complicated technical undertaking in order to realize this experiment. Bill Buxton is responsible for much of the conceptualization behind the design of the various systems. He also contributed to the content of this paper. Gordon Kurtenbach configured the hardware for the speech tracking equipment, and invested a great deal of time writing the software for recording, editing, and analyzing the speech time lines. Tom Milligan and Gary Hardock helped to set up and run the experiment, and helped troubleshoot when necessary. The patience of all the people in the CAVECAT project and the Dynamic Graphics Project at the University of Toronto is also much appreciated.

I also gratefully acknowledge the contribution of the Arnott Design Group of Toronto for the design and fabrication of the *Hydra* models. The work described in this paper has been supported by the Ontario Information Technology Research Centre, the Natural Sciences and Engineering Research Council of Canada, Xerox Palo Alto Research Center, Rank Xerox Cambridge EuroPARC, The Arnott Design Group (Toronto), Object Technology International (Ottawa), Digital Equipment Corp. (Maynard, MA.), and IBM Canada Laboratory Centre for Advanced Studies (Toronto).

REFERENCES

1. Argyle, M. (1975). *Bodily communication.* London: Methuen & Co. Ltd.

2. Argyle, M. and Cook, M. (1976). *Gaze and mutual gaze.* London: Cambridge University Press.

3. Argyle, M., Ingham, R., Alkena, F. and McCallin, M. (1973). The different functions of gaze. *Semiotica, 7,* 10-32.

4. Argyle, M., Lalljee, M., and Cook, M. (1968). The effects of visibility on interaction in a dyad. *Human Relations, 21,* 3-17.

5. Brady, P.T. (1968). A statistical analysis of on-off patterns in 16 conversations. *The Bell System Technical Journal, Jan.,* 73-91.

6. Cohen, K. M. (1982). Speaker interaction: Video teleconferences versus face-to-face meetings. *Proceedings of Teleconferencing and Electronic Communications,* University of Wisconsin, 189-199.

7. Cook, M. & Lalljee, M.G. (1972). Verbal substitutes for visual signals in interaction. *Semiotica, 3,* 212-221.

8. Dabbs, J.M. Jr., & Ruback, R.B. (1984). Vocal patterns in male and female groups. *Personality and Social Psychology Bulletin, 10(4),* 518-525.

9. Dabbs, J.M. Jr., & Ruback, R.B. (1987). Dimensions of group process: Amount and structure of vocal interaction. In *Advances in Experimental Social Psychology, 20,* 123-169.

10. Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology, 23(2),* 283-292.

11. Duncan, S. & Fiske, D. W. (1977). *Face-to-face interaction: Research methods and theory.* Hillsdale, NJ: Erlbaum.

12. Duncan, S. & Niederehe, G. (1974). On signalling that it's your turn to speak. *Journal of Experimental Social Psychology, 23,* 234-247.

13. Heath, C. & Luff, P. (1991). Disembodied conduct: Communication through video in a multi-media office environment. *Proceedings of CHI '91,* ACM Conference on Human Factors in Software, New Orleans, LA., 99-103.

14. Jaffe, J., & Feldstein, S. (1970). *Rhythms of dialogue.* New York: Academic Press.

15. Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 32,* 1-25.

16. Mantei, M., Baecker, R., Sellen, A., Buxton, W., Milligan, T. & Wellman, B. (1991). Experiences in the use of a media space. *Proceedings of CHI '91,* ACM Conference on Human Factors in Software, New Orleans, LA., 203-208.

17. Rutter (1987). *Communicating by telephone.* New York: Pergamon Press.

18. Rutter, D.R. & and Stephenson, G.M (1977). The role of visual communication in synchronizing conversation. *European Journal of Social Psychology, 2,* 29-37.

19. Sellen, A., Buxton, W. & Arnott, J. (1991). *Using spatial cues to improve desktop video conferencing.* 8 minute videotape. Toronto: Dynamic Graphics Project, Computer Systems Research Institute, University of Toronto.

20. Shannon, C.E., and Weaver, W. (1949). *The mathematical theory of communication.* Urbana, IL: University of Illinois Press.

21. Short, J., Williams, E., and Christie, B. (1976). *The social psychology of telecommunications.* London: Wiley & Sons.